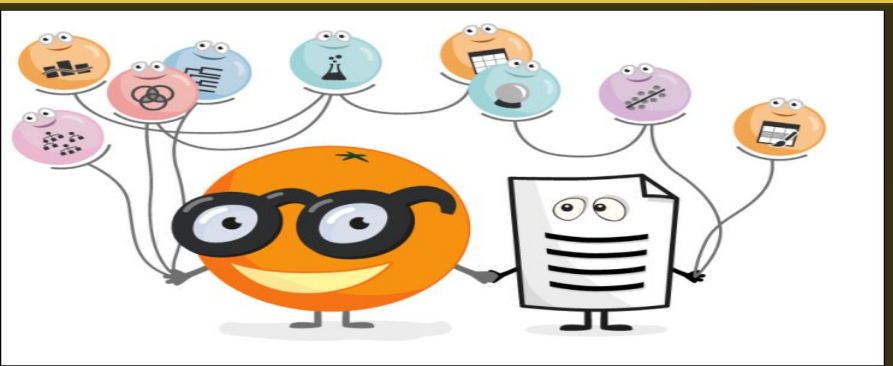
KOOKMIN UNIVERSITY

# 인공지능 데이터 분석
# orange

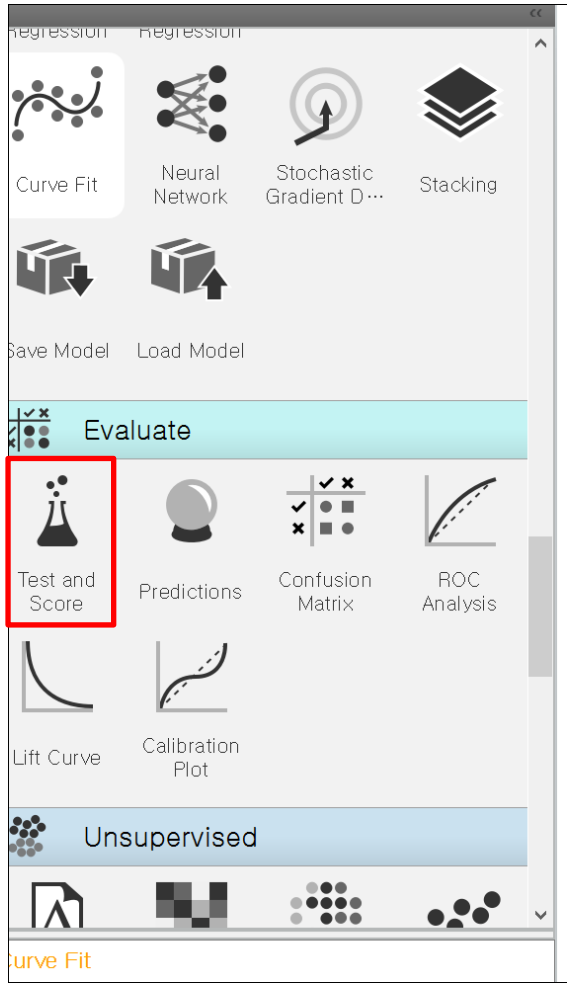


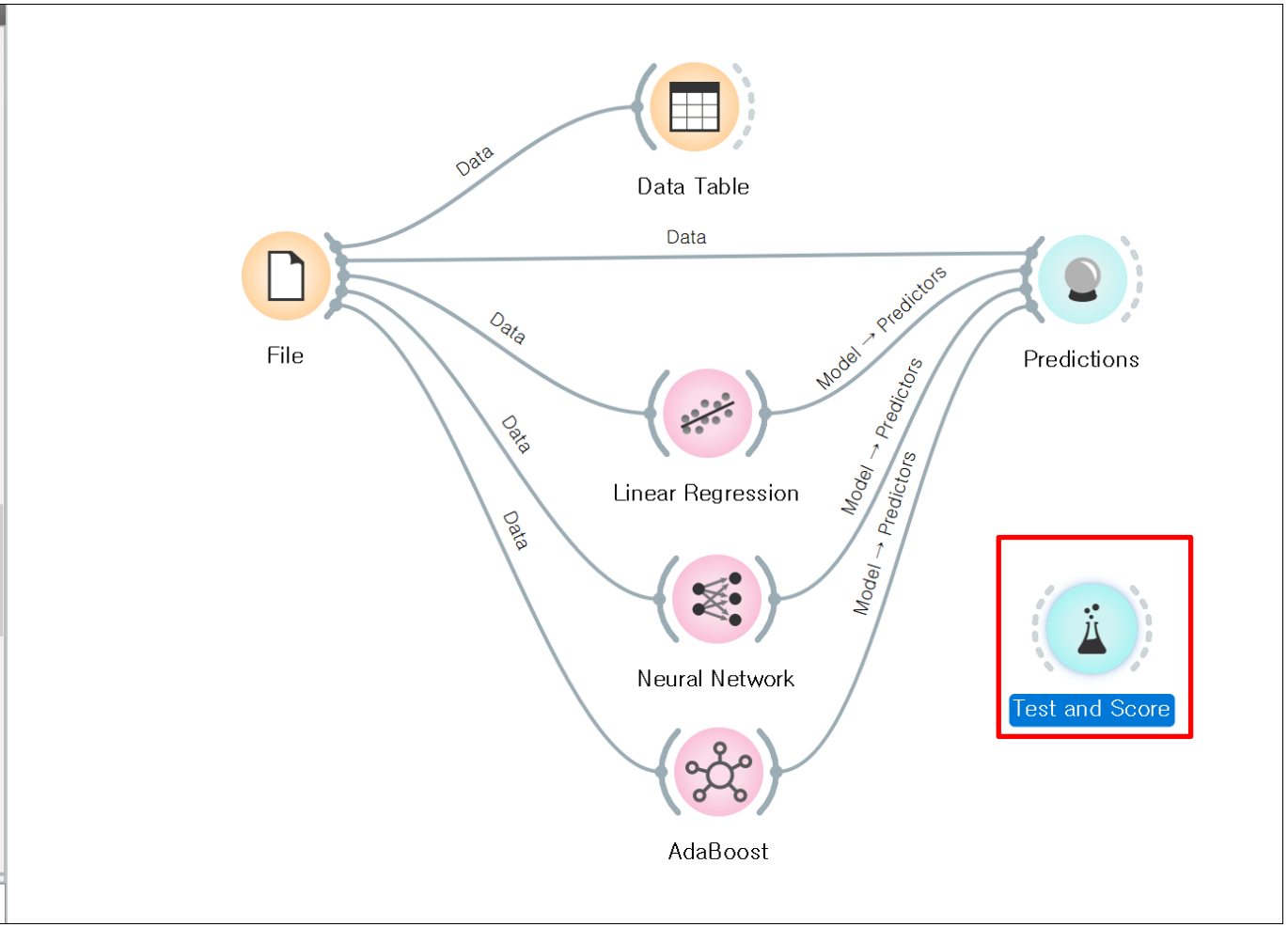
소프트웨어융합대학원
진혜진

## boston-housing-price

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | MEDV | CAT. MED |
| 2 | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 | 396.9 | 4.98 | 24 | 0 |
| 3 | 0.02731 | 0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 396.9 | 9.14 | 21.6 | 0 |
| 4 | 0.02729 | 0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 392.83 | 4.03 | 34.7 | 1 |
| 5 | 0.03237 | 0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 394.63 | 2.94 | 33.4 | 1 |
| 6 | 0.06905 | 0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 396.9 | 5.33 | 36.2 | 1 |
| 7 | 0.02985 | 0 | 2.18 | 0 | 0.458 | 6.43 | 58.7 | 6.0622 | 3 | 222 | 18.7 | 394.12 | 5.21 | 28.7 | 0 |
| 8 | 0.08829 | 12.5 | 7.87 | 0 | 0.524 | 6.012 | 66.6 | 5.5605 | 5 | 311 | 15.2 | 395.6 | 12.43 | 22.9 | 0 |
| 9 | 0.14455 | 12.5 | 7.87 | 0 | 0.524 | 6.172 | 96.1 | 5.9505 | 5 | 311 | 15.2 | 396.9 | 19.15 | 27.1 | 0 |
| 10 | 0.21124 | 12.5 | 7.87 | 0 | 0.524 | 5.631 | 100 | 6.0821 | 5 | 311 | 15.2 | 386.63 | 29.93 | 16.5 | 0 |
| 11 | 0.17004 | 12.5 | 7.87 | 0 | 0.524 | 6.004 | 85.9 | 6.5921 | 5 | 311 | 15.2 | 386.71 | 17.1 | 18.9 | 0 |
| 12 | 0.22489 | 12.5 | 7.87 | 0 | 0.524 | 6.377 | 94.3 | 6.3467 | 5 | 311 | 15.2 | 392.52 | 20.45 | 15 | 0 |
| 13 | 0.11747 | 12.5 | 7.87 | 0 | 0.524 | 6.009 | 82.9 | 6.2267 | 5 | 311 | 15.2 | 396.9 | 13.27 | 18.9 | 0 |
| 14 | 0.09378 | 12.5 | 7.87 | 0 | 0.524 | 5.889 | 39 | 5.4509 | 5 | 311 | 15.2 | 390.5 | 15.71 | 21.7 | 0 |
| 15 | 0.62976 | 0 | 8.14 | 0 | 0.538 | 5.949 | 61.8 | 4.7075 | 4 | 307 | 21 | 396.9 | 8.26 | 20.4 | 0 |
| 16 | 0.63796 | 0 | 8.14 | 0 | 0.538 | 6.096 | 84.5 | 4.4619 | 4 | 307 | 21 | 380.02 | 10.26 | 18.2 | 0 |
| 17 | 0.62739 | 0 | 8.14 | 0 | 0.538 | 5.834 | 56.5 | 4.4986 | 4 | 307 | 21 | 395.62 | 8.47 | 19.9 | 0 |
| 18 | 1.05393 | 0 | 8.14 | 0 | 0.538 | 5.935 | 29.3 | 4.4986 | 4 | 307 | 21 | 386.85 | 6.58 | 23.1 | 0 |

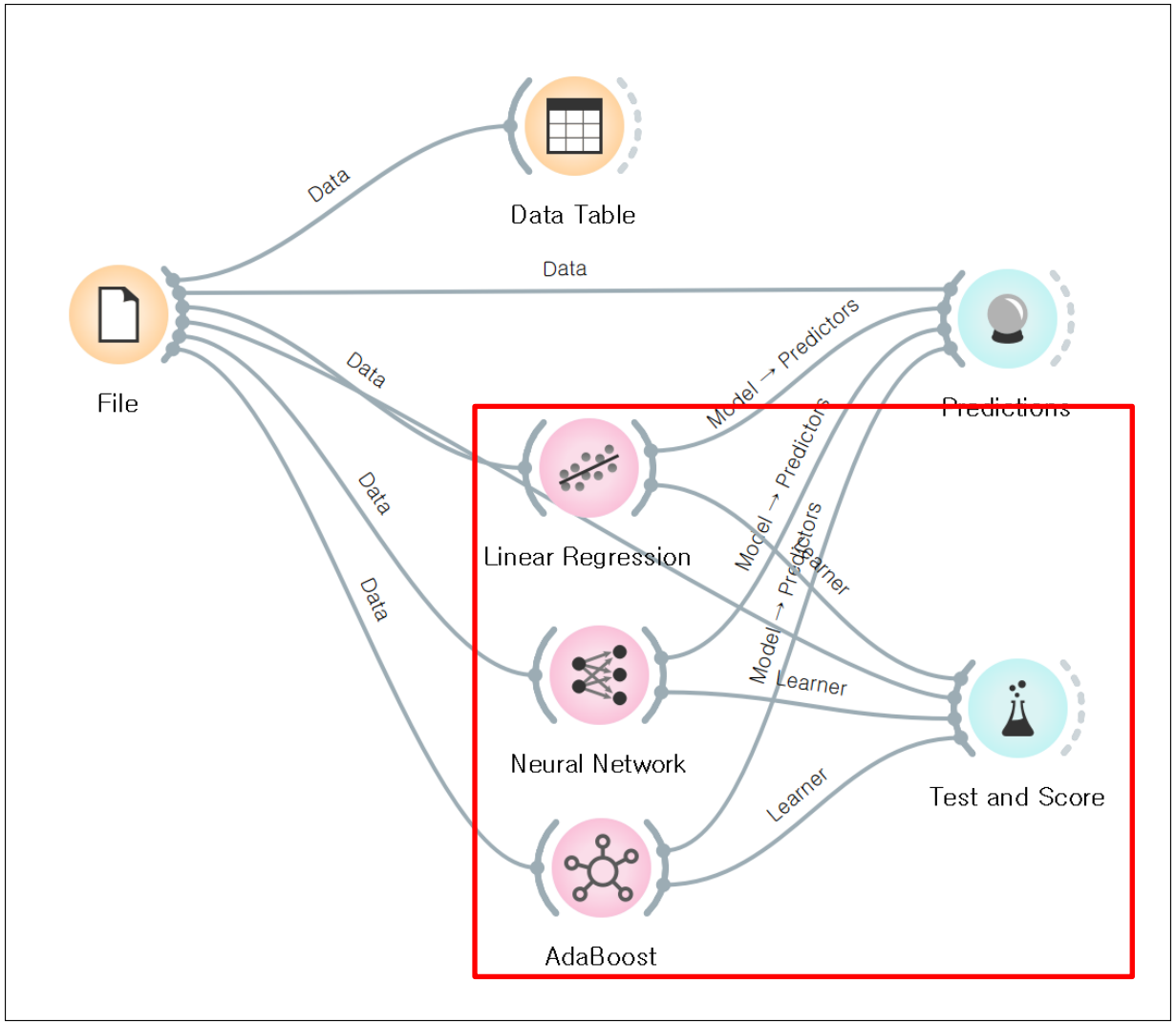
## Test and Score 위젯 추가



# 지도 학습_보스턴 주택 가격 데이터

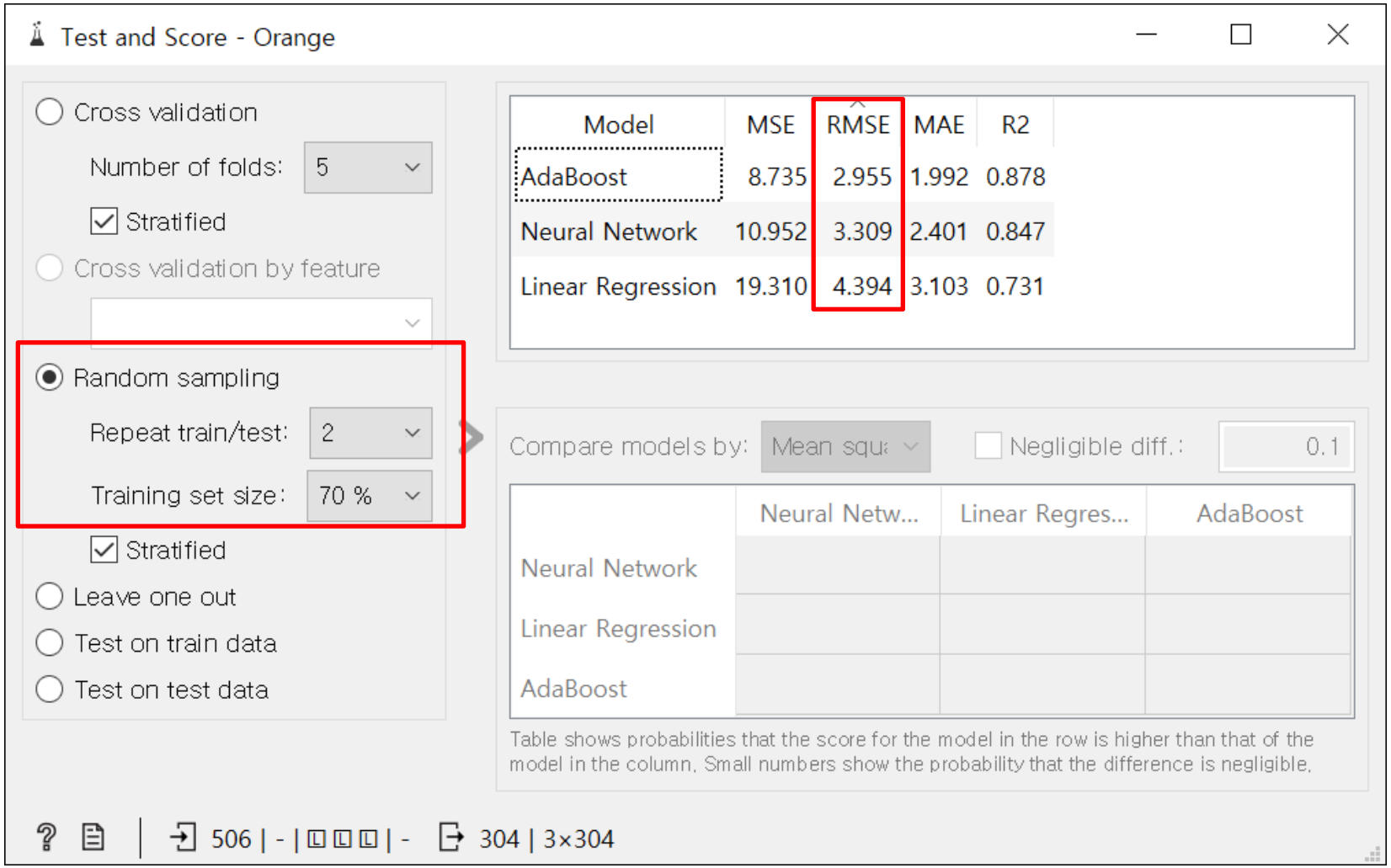




- **Random sampling**
  - Repeat train/test
    - 랜덤 샘플링을 몇 번 수행할 지 지정한다.
  - Training set size
    - 전체 데이터 중 학습을 위해 사용할 데이터의 비율이다.
    - 이 값을 70으로 지정하면 학습을 위해서는 70%를 쓰고 나머지 30%는 테스트 할 때 쓴다는 뜻이다.

# 지도 학습_보스턴 주택 가격 데이터

## Random sampling

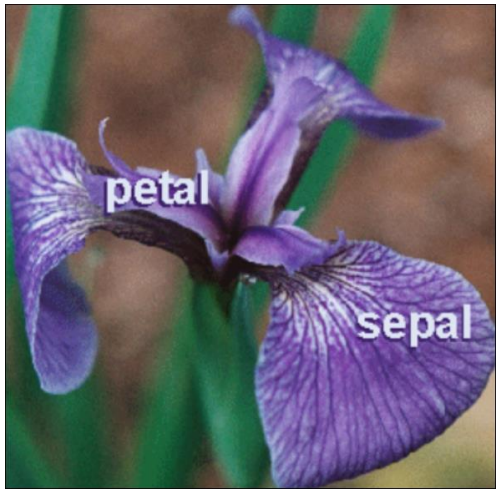


# 분류_붓꽃 데이터



- 붓꽃
  - 바깥쪽 면 꽃받침(sepal)
  - 안쪽 면 꽃잎(petal)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | sepal length | sepal width | petal length | petal width | iris |
| 2 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 3 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 4 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 5 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 6 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 7 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 8 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 9 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 10 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 11 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 12 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| 13 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 14 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa |
| 15 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa |
| 16 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa |
| 17 | 5.7 | 4.4 | 1.5 | 0.4 | Iris-setosa |
| 18 | 5.4 | 3.9 | 1.3 | 0.4 | Iris-setosa |
| 19 | 5.1 | 3.5 | 1.4 | 0.3 | Iris-setosa |
| 20 | 5.7 | 3.8 | 1.7 | 0.3 | Iris-setosa |
| 21 | 5.1 | 3.8 | 1.5 | 0.3 | Iris-setosa |
| 22 | 5.4 | 3.4 | 1.7 | 0.2 | Iris-setosa |
| 23 | 5.1 | 3.7 | 1.5 | 0.4 | Iris-setosa |
| 24 | 4.6 | 3.6 | 1 | 0.2 | Iris-setosa |
| 25 | 5.1 | 3.3 | 1.7 | 0.5 | Iris-setosa |
| 26 | 4.8 | 3.4 | 1.9 | 0.2 | Iris-setosa |
| 27 | 5 | 3 | 1.6 | 0.2 | Iris-setosa |



# 분류_붓꽃 데이터

## ■ 데이터 준비하기

File - Orange

Source

File: IRIS.csv

URL:

File Type

Automatically detect type

Info

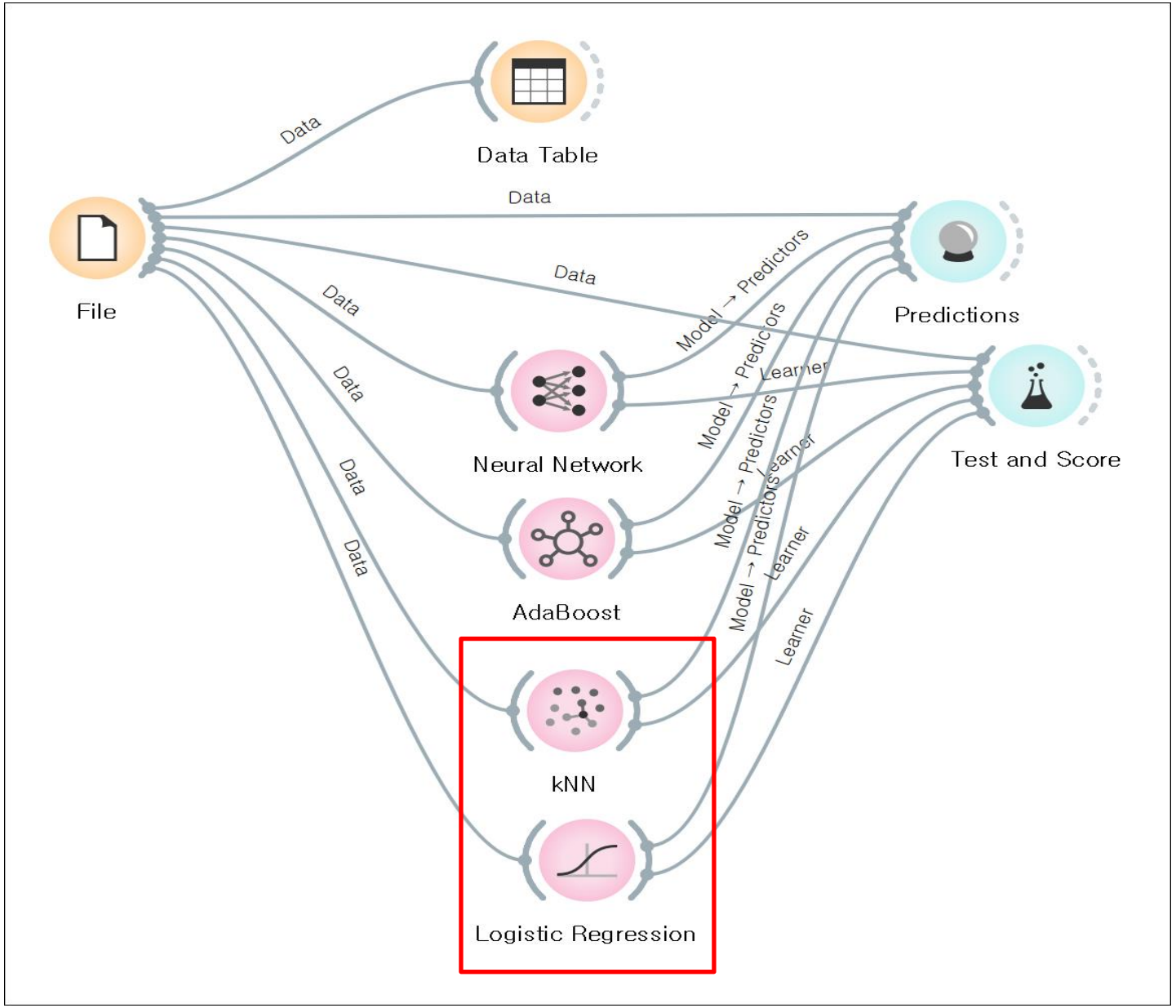
150 instance(s)
5 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

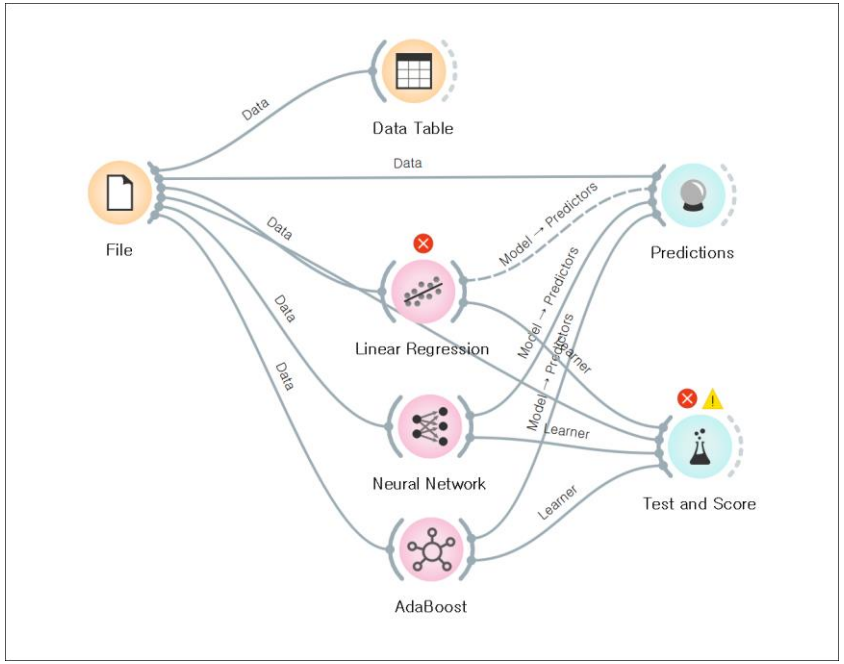
Columns (Double click to edit)

| | Name | Type | Role | Values |
|---|---|---|---|---|
| 1 | sepal length | numeric | feature | |
| 2 | sepal width | numeric | feature | |
| 3 | petal length | numeric | feature | |
| 4 | petal width | numeric | feature | |
| 5 | **iris** | **categorical** | **target** | **Iris-setosa, Iris-versicolor, Iris-virginica** |

# 분류_붓꽃 데이터




Data
Data Table
Data
File
Data
Data
Predictions
Model → Predictors
Data
Learner
Data
Model → Predictors
Neural Network
Test and Score
Data
Model → Predictors
Learner
Data
AdaBoost
Model → Predictors
Learner
Learner
kNN
Logistic Regression


Data
Data Table
Data
File
Data
Predictions
Model → Predictors
Linear Regression
Data
Model → Predictors
Learner
Data
Neural Network
Learner
Test and Score
Learner
AdaBoost

Predictions - Orange

Show probabilities for: Classes in data

Restore Original Order

| | Neural Network | AdaBoost | kNN | Logistic Regression | iris | sepal length | sepal width | petal length | petal width |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.97 : 0.03 : 0.00 → Iris-set... | Iris-setosa | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.99 : 0.01 : 0.00 → Iris-set... | Iris-setosa | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.99 : 0.01 : 0.00 → Iris-set... | Iris-setosa | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.97 : 0.03 : 0.00 → Iris-set... | Iris-setosa | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.99 : 0.01 : 0.00 → Iris-set... | Iris-setosa | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 5.0 | 3.4 | 1.5 | 0.2 |
| 9 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.97 : 0.03 : 0.00 → Iris-set... | Iris-setosa | 4.9 | 3.1 | 1.5 | 0.1 |
| 11 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 5.4 | 3.7 | 1.5 | 0.2 |
| 12 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.98 : 0.02 : 0.00 → Iris-set... | Iris-setosa | 4.8 | 3.4 | 1.6 | 0.2 |
| 13 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.97 : 0.03 : 0.00 → Iris-set... | Iris-setosa | 4.8 | 3.0 | 1.4 | 0.1 |
| 14 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.99 : 0.01 : 0.00 → Iris-set... | Iris-setosa | 4.3 | 3.0 | 1.1 | 0.1 |
| 15 | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 1.00 : 0.00 : 0.00 → Iris-set... | 0.99 : 0.01 : 0.00 → Iris-set... | Iris-setosa | 5.8 | 4.0 | 1.2 | 0.2 |
| 16 | 1.00 : 0.00 : 0.00 → Iris-set | 1.00 : 0.00 : 0.00 → Iris-set | 1.00 : 0.00 : 0.00 → Iris-set | 0.99 : 0.01 : 0.00 → Iris-set | Iris-setosa | 5.7 | 4.4 | 1.5 | 0.4 |

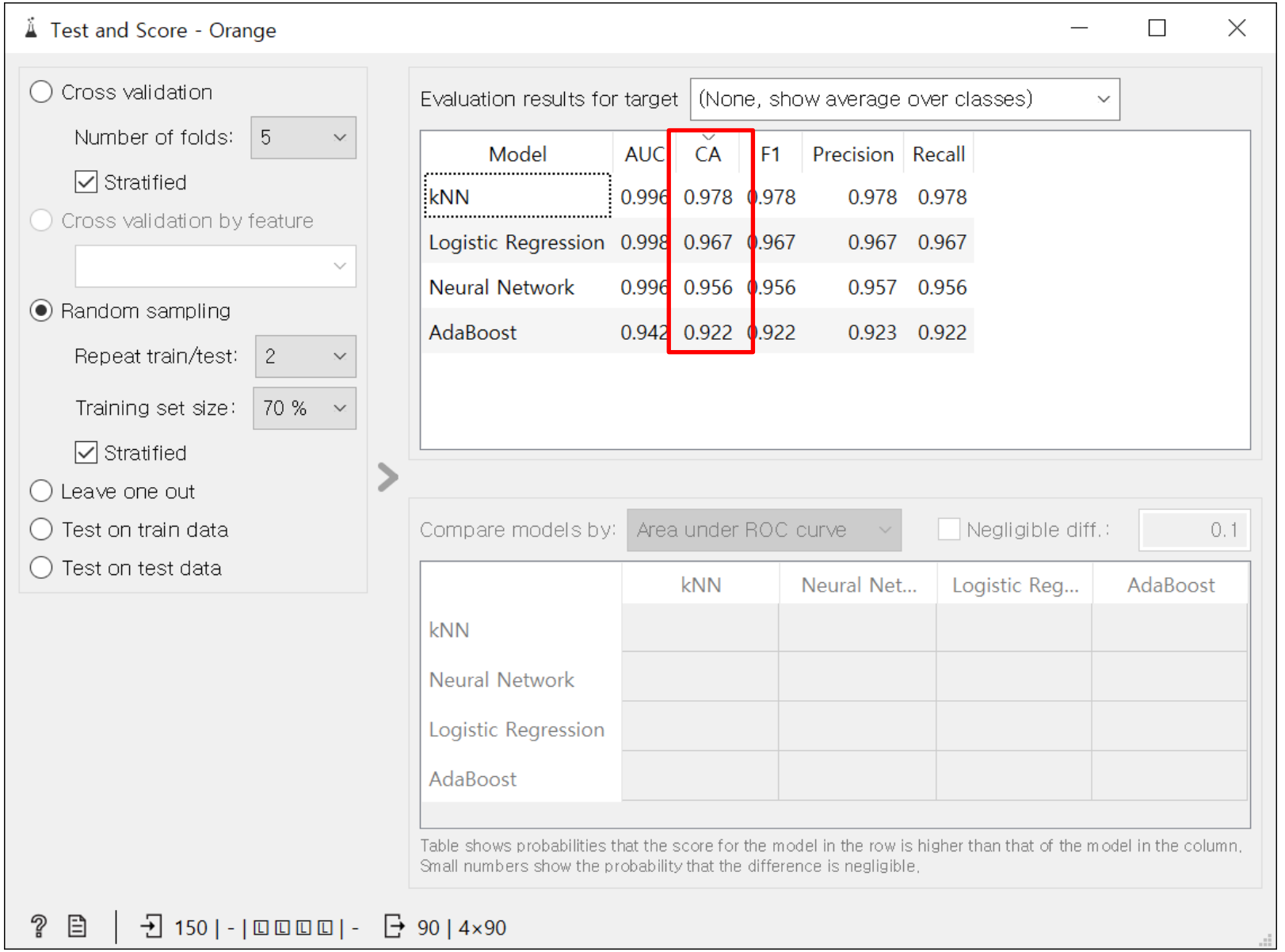
Show perfomance scores  Target class: (Average over classes)

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Neural Network | 0.999 | 0.987 | 0.987 | 0.987 | 0.987 |
| AdaBoost | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| kNN | 0.998 | 0.967 | 0.967 | 0.967 | 0.967 |
| Logistic Regression | 0.998 | 0.973 | 0.973 | 0.974 | 0.973 |

150 | M M M M  1 | 4×150

# 분류_붓꽃 데이터




Test and Score - Orange
Cross validation
Number of folds: 5
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 2
Training set size: 70 %
Stratified
Leave one out
Test on train data
Test on test data
Evaluation results for target (None, show average over classes)
Model AUC CA F1 Precision Recall
kNN 0.996 0.978 0.978 0.978 0.978
Logistic Regression 0.998 0.967 0.967 0.967 0.967
Neural Network 0.996 0.956 0.956 0.957 0.956
AdaBoost 0.942 0.922 0.922 0.923 0.922
Compare models by: Area under ROC curve
Negligible diff.: 0.1
kNN Neural Net... Logistic Reg... AdaBoost
kNN
Neural Network
Logistic Regression
AdaBoost
Table shows probabilities that the score for the model in the row is higher than that of the model in the column.
Small numbers show the probability that the difference is negligible.
150 | - | 90 | 4×90

- **좋은 모델을 찾는 방법**
  - 모든 학습 알고리즘을 다 이용해서 모델을 만들어보고, 모델끼리 경쟁시켜서 비교하는 것이다.
  - 아주 유명한 학습 방법 하나를 알아내서 언제나 그것만 쓰는 것이다.
  - 모든 학습 알고리즘을 깊이 있게 공부한 다음, 학습 시키고 싶은 데이터의 특성에 가장 잘 어울리는 학습 알고리즘으로 학습시키고, 그렇게 만들어진 모델을 경쟁시켜서 그중 가장 성능이 좋은 것을 사용하는 것이다.

## 지도학습 알고리즘 비교

Comparing Supervised Learning Algorithms : Table

| Algorithm | Problem Type | Results interpretable by you? | Easy to explain algorithm to others? | Average predictive accuracy | Training speed | Prediction speed |
|---|---|---|---|---|---|---|
| KNN | Either | Yes | Yes | Lower | Fast | Depends on n |
| Linear regression | Regression | Yes | Yes | Lower | Fast | Fast |
| Logistic regression | Classification | Somewhat | Somewhat | Lower | Fast | Fast |
| Naive Bayes | Classification | Somewhat | Somewhat | Lower | Fast (excluding feature extraction) | Fast |
| Decision trees | Either | Somewhat | Somewhat | Lower | Fast | Fast |
| Random Forests | Either | A little | No | Higher | Slow | Moderate |
| AdaBoost | Either | A little | No | Higher | Slow | Fast |
| Neural networks | Either | No | No | Higher | Slow | Fast |